Using electronic health records for clinical research: the case of the EHR4CR project

Georges de Moor, Mats Sundgren, Dipak Kalra, Andreas Schmidt, Martin Dugas, Brecht Claerhout, Töresin Karakoyun, Christian Ohmann, Pierre-Yves Lastic, Nadir Ammour, et al.

Using Electronic Health Records for Clinical Research: the Case of the EHR4CR Project

Georges De Moor[a,b,g], Mats Sundgren[c], Dipak Kalra[d], Andreas Schmidt[e], Martin Dugas[f], Brecht Claerhout[g], Töresin Karakoyun[h], Christian Ohmann[h], Pierre-Yves Lastic[i], Nadir Ammour[i], Rebecca Kush[j], Danielle Dupont[k], Marc Cuggia[l], Christel Daniel[m,n], Geert Thienpont[o], Pascal Coorevits[a,b]

[a] Department of Public Health, Unit of Medical Informatics and Statistics, Ghent University, Ghent, Belgium

[b] The European Institute for Health Records (EuroRec), Sint-Martens-Latem, Belgium

[c] AstraZeneca R&D, Mölndal, Sweden

[d] University College London, London, UK

[e] Pharma Product Development, F Hoffmann-La Roche Ltd, Basel, Switzerland

[f] Institute of Medical Informatics, University of Münster, Münster, Germany

[g] Custodix NV, Sint-Martens-Latem, Belgium

[h] Coordination Centre for Clinical Trials, Heinrich-Heine-University, Duesseldorf, Germany

[i] Sanofi R&D, Chilly-Mazarin, France

[j] Clinical Data Interchange Standards Consortium (CDISC)

[k] Data Mining International SA, Geneva, Switzerland

[l] INSERM CIC-P, CHU de Rennes, Rennes, France

[m] INSERM, UMR_S 1142, LIMICS, Paris, France

[n] CCS SI Patient, AP-HP, Paris, France

[o] Research in Advanced Medical Informatics and Telematics vzw (RAMIT), Ghent, Belgium

Corresponding author: Pascal Coorevits, Department of Public Health, Unit of Medical Informatics and Statistics, Ghent University, c/o Ghent University Hospital, Building K3 – 5th floor, De Pintelaan 185, 9000 Ghent, Belgium (tel +32 9 3328926, email: pascal.coorevits@telenet.be)

Keywords: clinical research, electronic health record, data reuse, clinical trials, interoperability, innovative medicines initiative, pharmaceutical industry, drug development

Abstract

Objectives: To describe the IMI EHR4CR project which is designing and developing, and aims to demonstrate, a scalable, widely acceptable and efficient approach to interoperability between EHR systems and clinical research systems.

Methods: The IMI EHR4CR project is combining and extending several previously isolated state-of-the-art technical components through a new approach to develop a platform for reusing EHR data to support medical research. This will be achieved through multiple but unified initiatives across different major disease areas (e.g. cardiovascular, cancer) and clinical research use cases (protocol feasibility, patient identification and recruitment, clinical trial execution and serious adverse event reporting), with various local and national stakeholders across several countries and therefore under various legal frameworks.

Results: An initial instance of the platform has been built, providing communication, security and terminology services to the eleven participating hospitals and ten pharmaceutical companies located in seven European countries. Proof-of-concept demonstrators have been built and evaluated for the protocol feasibility and patient recruitment scenarios. The specifications of the clinical trial execution and the adverse event reporting scenarios have been documented and reviewed.

Conclusions: Through a combination of a consortium that brings collectively many years of experience from previous relevant EU projects and of the global conduct of clinical trials, of an approach to ethics that engages many important stakeholders across Europe to ensure acceptability, of a robust iterative design methodology for the platform services that is anchored on requirements of an underlying Service Oriented Architecture that has been designed to be scalable and adaptable, EHR4CR could be well placed to deliver a sound, useful

and well accepted pan-European solution for the reuse of hospital EHR data to support clinical research studies.

1. Introduction

The rapid advancement and availability of health Information and Communication Technologies (ICT) offers remarkable enhancement opportunities for the clinical research sector [1]. Electronic health records (EHRs), interconnected through health care networks, have the potential to interact richly with research platforms [2]. The EHR in its ideal form for patient care is a longitudinal record of patient health information generated by multiple encounters in any care delivery setting. While the benefits of EHRs in direct patient care are widely recognised, deriving benefits from the reuse of EHR data through data warehousing for research purposes is still rare, underestimated or overlooked [3, 4]. In a survey of US academic health centres only 8% of respondents reported integration of clinical research data with patient clinical data [5]. If an EHR is fully implemented, reuse of EHR data may be extremely helpful in supporting clinical research by reducing redundant data capture, providing better understanding of real patient populations, supporting hypothesis testing, checking clinical trial feasibility, screening populations, supporting patient recruitment and early detection of safety risks, assessing treatment effectiveness and outcomes, and conducting post-marketing monitoring and long-term surveillance [6]. As an example, linking EHRs with clinical trials has proven to increase the recruitment rate of patients [7-9].

However, there are many obstacles to be overcome in using EHRs for clinical research. Fragmentation of patient records and proprietary health information technology systems that do not adhere to standards are a challenge. EHR vendors adopt few, if any, health information standards and very rarely accommodate controlled terminologies [10]. After evaluation of the numerous and varying initiatives across Europe, it is apparent that widespread incompatibility of the many data standards currently used by the clinical research and healthcare communities continues to hinder the efficient and rapid exchange of data between different electronic

sources and compromises the quality of clinical trial results. Additional challenges that currently limit the use of, and value derived from, health ICT solutions in Europe include regional diversity in languages, healthcare practices and regulations, the emergence of multiple non-interoperable hospital EHR systems, and inadequate and inconsistent clinical documentation within EHRs. These limitations currently prevent the optimal use of EHR patient level data and information, and impede the advancement of medical research, the improvement of healthcare and the enhancement of patient safety [11].

Processes and technologies that meet data privacy and regulatory requirements, and satisfy other organisational governance stipulations, are necessary prerequisites to gaining acceptance of the reuse of EHRs for research [12, 13]. Privacy, legal implications and public relations ramifications were stated as concerns by over 80% in a survey [14]. Data quality (consistency, correctness, completeness) is another challenge for the reuse of EHR data [15-17]. As a consequence, EHRs appear to mainly be used currently for non-regulated investigator-sponsored clinical research and evaluating pilot or prototype applications [7, 10, 14, 18].

Moving beyond the current state-of-the-art implies setting up a framework to enable interoperability (encompassing technical, standards, functional, legal, and organisational aspects) for future electronic data collection and exchange between systems. This paper presents the current results of the IMI EHR4CR project which has designed, developed, and aims to demonstrate, a scalable, widely acceptable and efficient approach to interoperability between EHR systems and clinical research systems.

2. Background

2.1 Overview of the IMI program

The Innovative Medicine Initiative (IMI) was launched in 2008 by the European Union and the European Federation of Pharmaceutical Industries and Associations (EFPIA), with a total budget of 2 billion Euros to be spent over a 10-year period, making IMI the largest public-private partnership (PPP) in life sciences R&D. IMI's two founding members, EFPIA and the European Union, have equal investment and rights in the IMI. To fulfil its mission, the IMI implements R&D programs focused on developing new tools and methods for predicting drug safety and efficacy as well as for more efficient knowledge management. EFPIA pharmaceutical companies invest in the IMI in the form of in-kind contributions by committing internal human resources or providing access to data sets and infrastructure and sometimes in the form of direct monetary contributions. This industry investment is matched by funds from the European Union; the funds support other consortium members, including academic teams, small and medium-sized enterprises (SMEs), patient organisations, regulatory agencies, and relevant not-for-profit institutions. The first three calls for proposals, launched by the IMI in 2008, 2009, and 2011, resulted in 30 projects including EHR4CR.

2.2 Overview of the EHR4CR Project

The EHR4CR project runs over 5 years (2011-2015) with a budget of +16 million Euros, involves 34 academic and private partners (10 pharmaceutical companies) and is to date one of the largest of the IMI PPPs in this area [19]. This project is developing adaptable, reusable and scalable solutions (tools and services) for reusing data from EHR systems for clinical research

purposes. The consortium also includes 11 hospital sites in France, Germany, Poland, Switzerland and the United Kingdom.

3. Methods

3.1 General overview

The EHR4CR project aims to build a robust and scalable platform that will unlock data from hospital EHR systems, in full compliance with the ethical, regulatory and data protection policies and requirements of each participating country. The EHR4CR platform supports distributed querying to assist in clinical trials feasibility assessment and patient recruitment, and will in a later version also link EHR systems and EHR4CR services to enable clinical researchers to obtain key information about a patient's health and healthcare history before they arrive for a screening visit (but after patient consent has been obtained). Such developments require securing acceptance from the patients, the public and the research and health service communities. Therefore, in parallel to the technical developments, senior level decision makers, ethics boards and industry executives and scientists, are involved in consultations to provide strategic insights into the most robust and acceptable technical and procedural approaches that should be taken to ensure privacy protection and compliance with European and national/regional regulations on data protection.

The platform is currently being piloted at several hospital sites across Europe. These sites are themselves active in clinical research and are able to provide exemplary local governance requirements to complement the ethical inputs referred to above. To enable wide adoption by EHR vendors and quality assurance of the EHR4CR platform within hospitals, the project also

provides governance through accreditation/certification programs for establishing best practices.

3.2 Stakeholders engagement

Several tasks are being deployed to research and develop stakeholder and ethical-legal requirements for the EHR4CR platform and clinical research services. The requirements engineering process is a sub-process as well as a sub-part of the iterative and incremental development strategy of the EHR4CR project. The basic idea is to develop a system through repeated cycles. Starting with a subset of the software requirements, the system is being iterated until the full EHR4CR platform is specified. In each iteration, design modifications are made and new functional capabilities are added. In the first step we defined for each iteration one domain scenario that is used for estimating the probable effects, as an integral part of situation analysis and long-range planning and describes the entire domain, e.g. protocol feasibility. In the next step the domain scenario has been broken down into high-level 'Usage Scenarios' that describe the critical business interactions (the goals, motivations, inputs, steps, events, and/or actions which occur during the interaction) with enough detail to indicate their anticipated operation for the delivery, control and use of clinical research services. The usage scenarios serve as the context for the use cases and requirements and allow the teams to make sure they are complete. Evaluation criteria have also been defined.

3.3 Technology platform and tools

Based on an inventory of relevant solutions and components external to the project as well as the internal requirement specifications, a reference architecture was defined to serve as a

technical specification for the construction of a scalable platform supporting the EHR4CR clinical research services. This EHR4CR platform is implemented as a common set of components and services that will allow the integration of the lifecycle of clinical studies with heterogeneous clinical systems, hereby facilitating data extraction and aggregation, workflow interactions, privacy protection, information security, and compliance with ethical, legal and regulatory requirements. This will help speed up the protocol feasibility refinement process with rapid feedback on population numbers and their geographic distribution, assist in identifying suitable patients via their nominated care providers, speed up and improve the accuracy of patient recruitment and trial execution, and enable more complete and real time safety monitoring.

Tools and services are provided to ensure interoperability between varying and disparate data sources (EHR and EDC Systems), allowing for the consistent interpretation of data available from those sources by the EHR4CR end-user services: a key success factor for the project. A central EHR4CR "pivot terminology", which is largely based on existing standards such as ICD10, LOINC, and SNOMED-CT, acts as the principal broker between heterogeneous terminology systems and is used to maintain mappings from central (platform-level) concepts to local (data source-level) concepts and vice-versa [20,21]. The necessary software applications are being developed to deliver the EHR4CR clinical research services to end users: clinical protocol feasibility, patient identification and recruitment, clinical trial execution, and severe adverse event reporting.

3.4 Pilots

The EHR4CR platform is being evaluated by demonstrating the functionality of the tools and services. These evaluations occur at several large academic hospitals, interfacing with EHR

systems, with a specific focus towards a set of medical domains mutually agreed between the pilot sites and EFPIA partners. The EHR4CR project primarily addresses the following disease areas included in the pilots: oncology, inflammatory diseases, neuroscience, diabetes, cardiovascular and respiratory diseases. These disease areas are relevant to current pharmaceutical industry pipelines, and align with clinical research interests and data resources at the pilot sites. Interfaces between the EHR systems and the central EHR4CR platform have been established. An inventory of data elements for pilot studies has been defined. Semantic mapping between local terminologies and the central EHR4CR terminology has been conducted. Clinical data warehouses (CDWs), compliant with the EHR4CR platform and the associated extract-transform-load (ETL) processes have been designed and tested. Approval of all data processing steps was gained in accordance with local ethical and legal regulations at each site.

4. Results and discussion

4.1 Stakeholders engagement

A detailed analysis has been undertaken of the requirements for supporting protocol feasibility, patient identification and recruitment. The starting point for this was a study through interviews of protocol managers in order to understand their current workflows, and obstacles in undertaking these activities, focusing in particular on those that contribute to delays in undertaking clinical trials. Idealised usage scenarios that would take full advantage of a research platform were then formally documented and used to develop use cases. As a result of the requirements engineering process, six usage scenarios for protocol feasibility, 75 use cases and about 200 requirements were identified and delivered in a software requirement specification (SRS) document for the EHR4CR Protocol Feasibility Service. This SRS includes functional and non-functional requirements. Materials of a similar scale have now been completed for the EHR4CR Patient Identification and Recruitment Service. These requirements are now informing the evaluation criteria, which are being developed in collaboration with the pilot sites.

Early in the project a European electronic survey was undertaken to better understand the market landscape and help design a sustainable EHR4CR business model and value propositions. The survey was conducted in two waves via an online questionnaire i) with participating stakeholders from the public and private sectors involved in the EHR4CR consortium, and ii) with non-participating informed stakeholders. The results were highly consistent between the two groups and confirmed a high level of interest in the EHR4CR objectives and scenarios, and the relevance of developing customised value propositions to address the respective needs of key stakeholders. A scientific manuscript presenting the EHR4CR e-survey objectives, methods, results and conclusions has been published [22].

In order to understand the attitudes and concerns of senior level stakeholders across Europe to the approaches being developed in EHR4CR, and therefore to ensure their later support for a Europe wide EHR4CR deployment, an in-depth survey questionnaire was developed, with two variants for ethics and non-ethics stakeholders. The questionnaires were developed and initially piloted with 16 interviews undertaken in Scotland, providing the project with some early and important insights, prior to conducting these as in depth interviews in five other European countries. The detailed results of these interviews will be published later.

Internal EHR4CR policy documents and other written materials that may be furnished to hospital data protection officers have been developed to support the pilot sites in gaining local approval for connecting to the prototype platform and for providing a de-identified sample of patient data to validate the research query processes.

4.2 Technical platform and tools

An inventory of reusable and available relevant solutions and components has been compiled with respect to the functional and non-functional aspects. Examples include middleware, security tools, query engines and end-user applications developed by consortium partners in previous projects such as FARSITE [23] and ePRCN [24] as well as data querying and analysis tools with open source license models like i2b2 [25] and SHRINE [26] that are actively maintained by other organisations or communities. As an outcome of this effort, software libraries and methodologies have been re-used and software connectors have been provided (e.g. offering the possibility to use the EHR4CR reference implementation with an i2b2 clinical data warehouse).

4.2.1 Information representation

In addition, an inventory of information and knowledge models has been developed, including terminologies for patient care coordination, information models provided by standards development organisations for patient care coordination and information models available from other projects/initiatives.

In the recent past several projects, including eMERGE [27], i2b2-SHRINE [28], SHARPn [29], OHDSI [30] and the Office of the National Coordinator (ONC) Standards and Interoperability (S&I) Framework [31,32] and - pilots [33] have developed tools and technologies that bring out the value of observational health data through large-scale analytics for decision support and research. The developed solutions allow identifying patient cohorts and extracting patient-centric data using distributed EHRs/CDWs for defined purposes, including case–control cohorts for genome-wide association studies or pharmacogenetic studies, feasibility studies and patient recruitment in clinical trials, adverse event detection or quality metrics calculation.

In this context, “phenotyping algorithms” are defined in order to compute eligibility criteria for identifying patient cohorts and to collect patient-centric data of interest. “Phenotyping algorithms” are typically represented as pseudocode with varying degrees of formality and structure [29, 34].

The algorithmic patient cohort identification and data extraction within EHRs/CDWs is based on query specification involving EHR/CDW data fields (e.g., diagnoses, procedures, laboratory results, and medications) and logical operators. Clinical statements in the query specification expressed in the clinical research Model of Use first need to be defined using Common Data Elements of the Model of Meaning and then transformed into local-EHR-specific Model of Use to be executed within EHRs/CDWs.

A semi-formal representation of eligibility criteria has been developed which was used to identify a core set of data elements that cover an important part of patient-level inclusion and exclusion criteria. The EHR4CR semantic resources integrate a range of clinical models linked to clinical terminologies that are needed to collectively represent the variety of clinical statements (including diagnosis, findings, familial and medical history, lab tests, medication, etc). The layering of the EHR4CR information model is based on i) a generic reference information model i.e. the ISO/HL7 21731 Version 3 Reference Information Model (RIM), ii) more detailed information models i.e. the Common Element Templates (CETs) and Common Data Elements (CDEs) specified consistently with ISO/IEC 11179 Metadata Registry (MDR) standard [35] and with the ISO 21090 Healthcare Data Types [36] and iii) clinical terminology models such as ICD or SNOMED-CT.

The set of EHR4CR Common Element Templates (CETs) and Common Data Elements (CDEs) that instantiate generic reference models and are tailored to the needs of structured data acquisition (e.g. HL7's Clinical Document Architecture (CDA) meta-standard and the derived Continuity of Care Document (CCD) and Consolidated Clinical Document Architecture (C-CDA)).

EHR4CR Common Element Templates (CETs) and Common Data Elements (CDEs) can be considered as the semantic building blocks of the EHR4CR Common Information Model, of which each interoperating applications should be aware. EHR4CR Semantic Resource Repository provides to EHR4CR components and interoperating applications machine-processable definitions of CETs and CDEs that they can consume through semantic services during both query specification at workbench and query execution at endpoint. Semantic services are essential, and were developed according Common Terminology Services 2 (CTS2) functional specification that needed to be extended.

Semantic annotation of these data elements (and of the concepts of their value sets) has been undertaken based on reference terminologies integrated in the EHR4CR pivot terminology. The first version of the HL7 RIM-based EHR4CR information model (a platform-independent conceptual model) is now complete and its relationship with other reference information models such as CDISC/HL7 BRIDG (Biomedical Research Integrated Domain Group), a domain analysis model for the clinical research domain serving to link healthcare and research [37] or Observational Medical Outcomes Partnership (OMOP) Common Data Model [38], a standard format of observational data accessed in the context of drug safety analysis, are currently documented.

The EHR4CR semantic models, resources & semantic tools are available at [39].

4.2.2 Platform services

An EHR4CR platform architecture description has been developed and has been iteratively refined and extended. This has ensured that important concerns have been addressed by the architecture and has established a common understanding amongst the EHR4CR development teams. The platform is based on a Service Oriented Architecture (SOA) into which service providers and consumers can dynamically connect. As such, the primary goal of the EHR4CR architecture description is the specification of clearly defined interfaces and component responsibilities while the physical location of service consumers and providers is only of secondary importance. The key EHR4CR services and their behaviour are described in Table 1.

Service	Category	Description
Registry service	Infrastructure	Allows publishing and inquiring the technical capabilities and organization-level metadata of a (data provider) site exposing applications or services on the platform.
SSO Service	Security	Provides single-sign on for end-users to platform applications
Security Token Service	Security	Authenticates clients of platform (Web) Services. Also supports credential delegation allowing applications to act as end-users, thus allowing them to invoke services on their behalf with full traceability.
Identity Management Service	Security	Allows managing platform users and their organisations.
Message broker	Infrastructure	Allows data providers to interact with platform applications without having to expose their Web Services on the Internet.
Terminology Service	Semantic interoperability	Allows querying for central information model concepts (data elements, terminologies and value sets) and their relationships. Used for query specification at Workbench (accessing central codes from reference terminologies to specify canonical queries). Used for query execution at endpoint (transcoding central to local codes in queries and vice-versa in results)

Terminology Mapping Service	Semantic interoperability	Allows mapping central terminology elements and value sets to their local counterpart(s) and vice-versa.
Eligibility criteria query service	Querying	Allows executing protocol feasibility queries and retrieving their results.
Patient recruitment service	Workflow	Web Service exposed by a sponsor for receiving a clinical site's study participation status and recruitment status updates.
Patient recruitment participation service	Workflow	Web Service exposed by a clinical site for receiving study participation requests and study status updates.

Table 1: Key EHR4CR services

Data endpoints are key services in the EHR4CR platform from which the different scenarios can be built up. Given their importance and complexity, they are a particular focus of the EHR4CR architecture description. Figure 1 provides a simplified view of the EHR4CR platform, focusing on the data endpoints.

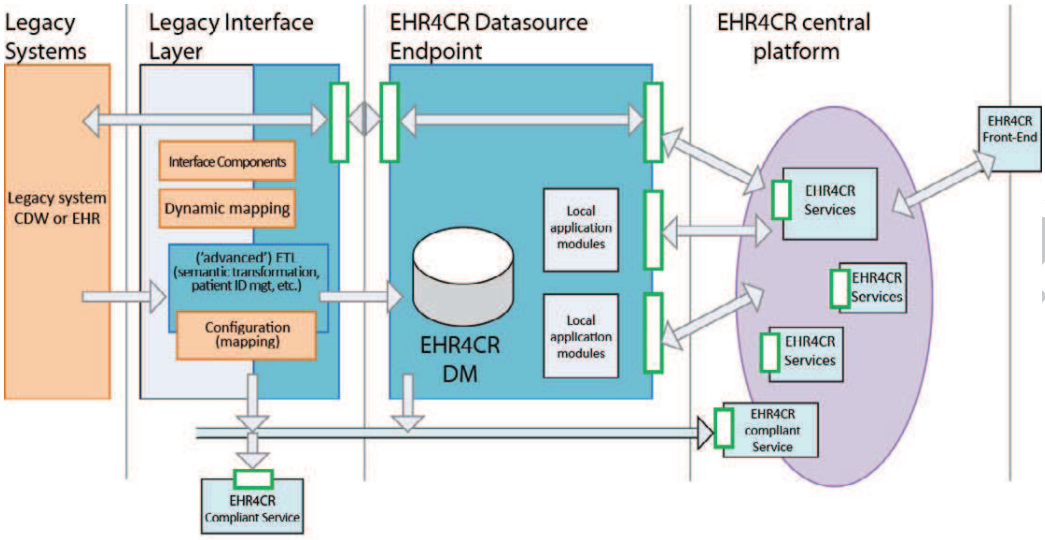


Figure 1: Simplified view of the EHR4CR platform

4.2.3 Data endpoint (hospital) services

The access to the clinical data locally at the data endpoints consists of three logical layers: the legacy system layer (specific to the type of CDW or EHR used by the local site), the legacy interface layer (for accessing the local CDW or EHR) and the EHR4CR Data source Endpoint layer (generic layer used by all the EHR4CR compatible data endpoints). The Legacy Interface layer deals with the complexity involved when accessing the various types of CDW and EHR systems used and the adopted strategy for translating queries against the EHR4CR information model and pivot terminology into queries that can be executed locally against the CDW or EHR system. The EHR4CR Data source Endpoint layer is a generic layer that exposes uniform EHR4CR endpoint interfaces to other EHR4CR services and components and allows for plugging in various local application modules for the different functional scenarios tackled within EHR4CR. The EHR4CR Data source Endpoint layer encompasses a generic EHR4CR Data Mart (DM) component that is used in case an ETL strategy has been adopted within the Legacy

Interface Layer. The EHR4CR DM is a direct physical representation of the EHR4CR information model. These layers can interact directly with other EHR4CR compliant services if required (e.g. invoking a measurement conversion service) (Figure 1). The EHR4CR Data source Endpoint reference implementation that has been developed in the EHR4CR project assumes an underlying SQL-based data warehouse and uses a system of query templates for translating individual criteria to queries against the local physical information model. It relies on a terminology mapping service for dynamically translating between concepts expressed using the pivot terminology and local terminologies. It contains capabilities for converting to and from UCUM measurements or for converting coded concepts to and from their corresponding rank in stages or scores (e.g. ECOG scores or TNM stages).

In order to be able to interact with other platform users and services, service providers (and in particular data providers) must adhere to the technical interfaces and support information exchange based on the information models. Additionally, they may need to support service-provider specific requirements. Once approved, the service provider metadata is added to the central registry and exposed services and applications can be published so that platform users can discover them.

4.2.4 Platform architecture for PFS and PRS

The current architecture description focuses on the Protocol Feasibility Scenario (PFS) and Patient identification and Recruitment Scenario (PRS). The main components involved in the protocol feasibility scenario are (see Figure 2):

- Protocol feasibility tools in the form of a workbench for studying non-identifiable distributed patient data. Note that these tools focus on authoring and managing

(computable) eligibility criteria queries rather than providing functionality for clinical trial protocol authoring;

- An orchestration module allowing distributed execution of eligibility criteria queries;
- Endpoint (data access) services allowing eligibility criteria query execution on local clinical data warehouse facilities;
- Supporting semantic interoperability services (e.g. coding system value mapping), registry services (e.g. for dynamically discovering query endpoints), a message broker and security services (e.g. *single sign-on*).

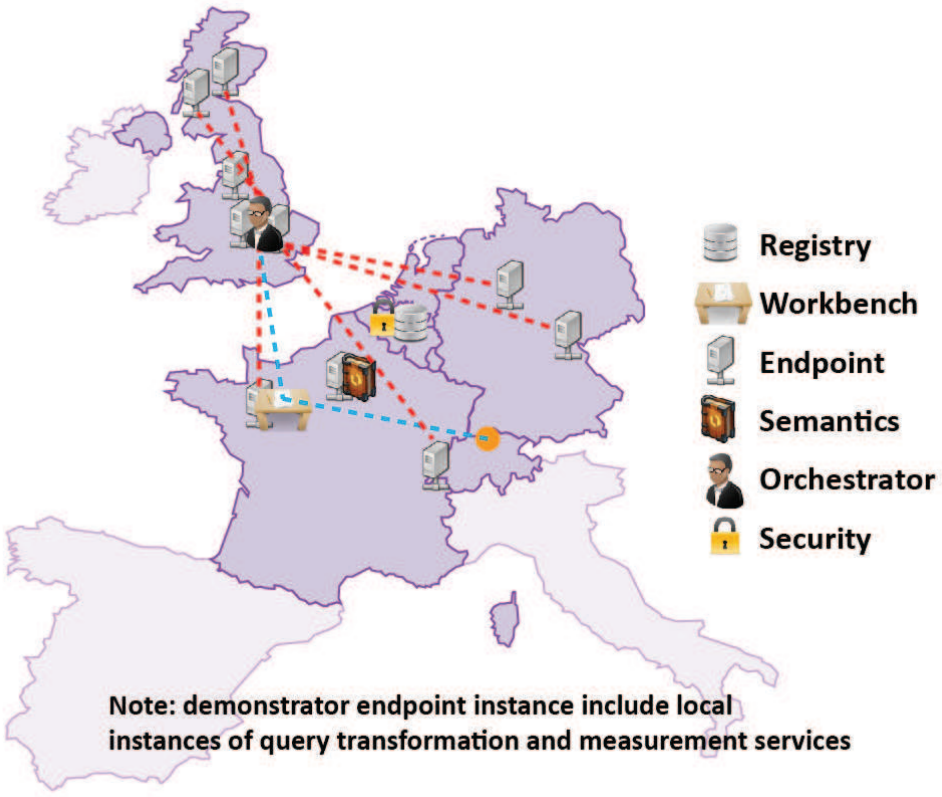


Figure 2: Overview of the main services of the Protocol Feasibility Scenario

In the PFS scenario, EHR4CR avoids issues related to patient informed consent as no individual patient-level information is being exchanged (i.e. leaves the hospital site), only aggregated numbers according to demographic categories. Additional measures such as data perturbation and generalization have been built in to avoid potential patient re-identification from aggregated results.

For the patient identification and recruitment scenario, the EHR4CR platform has been extended in the following ways:

- The workbench designed and developed for the PFS has been extended to support coordination and monitoring of the recruitment process.

- New patient identification and recruitment tracking tools have been made available to the clinical sites in the form of a workbench application for relationship management with clinical trial sponsors, local study management and patient identification and recruitment status tracking. Each participating clinical site has its own installation only to be used locally. Supporting endpoint (data access) services allow eligibility criteria query execution for composing the initial set of candidate patients for recruitment (to be assessed by a human investigator). Re-identification services locally allow authorised treating physicians to see the patient identifying information corresponding to a candidate record in order to allow getting into contact with the patient to initiate the recruitment process.

- Secured web services at both ends (local and central workbenches) allow exchanging the necessary information (such as accrual rates, but not patient identifiable

information) to steer the recruitment processes at the sponsor site and each of the individual clinical sites.

Similarly to PFS, PRS avoids patient consent-related issues by not exposing patient-level information outside the environment of the clinical sites (Figure 3). On the local level (within each clinical site), the PRS services support the most stringent requirements encountered amongst the consortium's pilot sites, requiring for instance that a treating physician (or an equivalent role by configuration) notifies each patient who is a potential candidate for recruitment and obtains the patient's approval before handing over the patient's personally identifiable information to a research investigator for further follow-up. The solution is based on a configurable workflow system that allows for customisation according to local rules and policies. For example, the assignment of the person or group to initially contact a potential candidate patient can be based on departmental information recorded in the clinical data warehouse or other information sources. In addition, tracking of obtained informed consent is an explicit step in the patient's recruitment status tracking system.

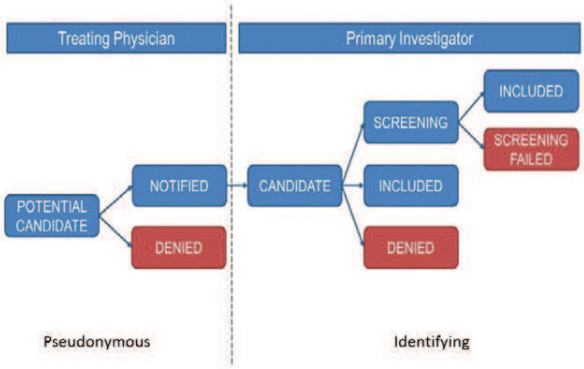


Figure 3: Patient recruitment status tracking and privacy

Viability of the initial EHR4CR architecture has been shown in the form of a PFS demonstrator incorporating reference implementations of the various platform tools and services described by the architecture and data endpoint instances hosted by the EHR4CR partners contributing to the pilot activities (see below).

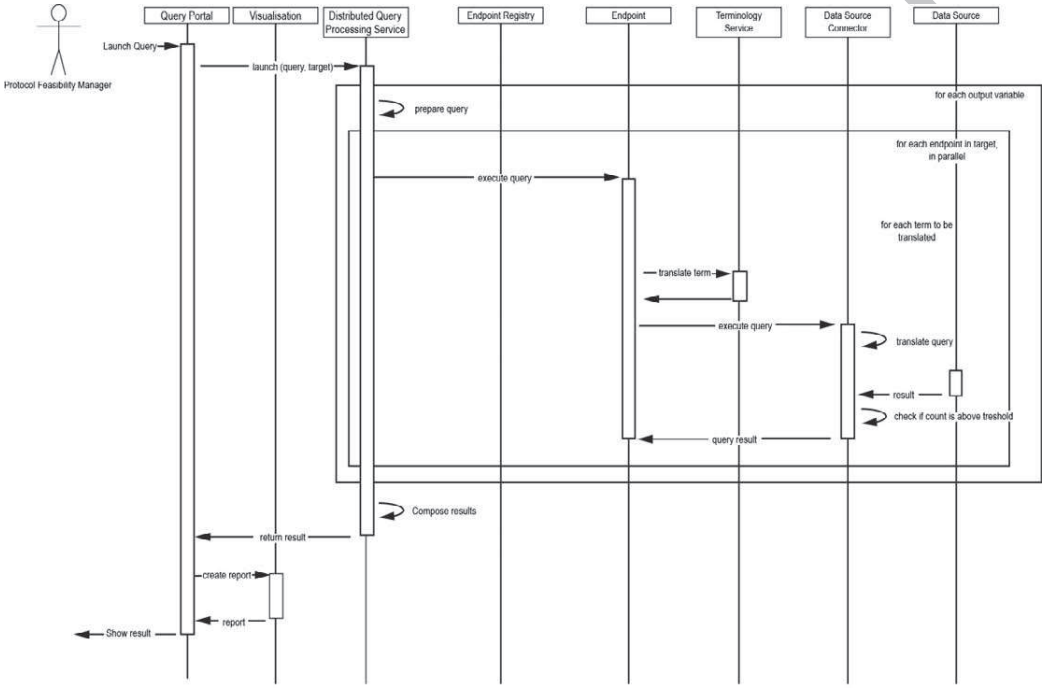


Figure 4: Sequence diagram for PFS query execution

Figure 2 illustrates the physical location of the main services involved in the PFS demonstrator while Figure 4 shows the main interactions. The process of querying individual data endpoints for protocol feasibility is initiated by the workbench instance on behalf of an end-user. The workbench then submits a series of eligibility criteria (EC) queries to an orchestrator service instance. The orchestrator instance identifies and invokes the data endpoints to which the EC

queries are targeted. Finally after receiving the individual EC query results (number of eligible patients at each site), the orchestrator service instance provides a consolidated result to the workbench instance which will eventually be displayed to the end-user. Discovery of the involved endpoints and their location is provided by a set of registry services. Transformation between EC against the EHR4CR information model and terminologies and local models and terminologies is provided by a number of semantic services. Finally, a number of platform-level authentication and authorisation services provide a uniform security model and ensure compliance with the EHR4CR security architecture (Figure 2).

4.2.5 Clinical research workbench and workflow

The PFS demonstrator includes a workbench application that allows the authoring and execution of computable Eligibility Criteria (EC) queries and allows secured sharing of feasibility studies and the associated EC queries amongst different platform users. EC queries can be built using a user-friendly graphical user interface which allows specifying Boolean and temporal constraints between individual EC (Figure 5) (see also [40] for more details).

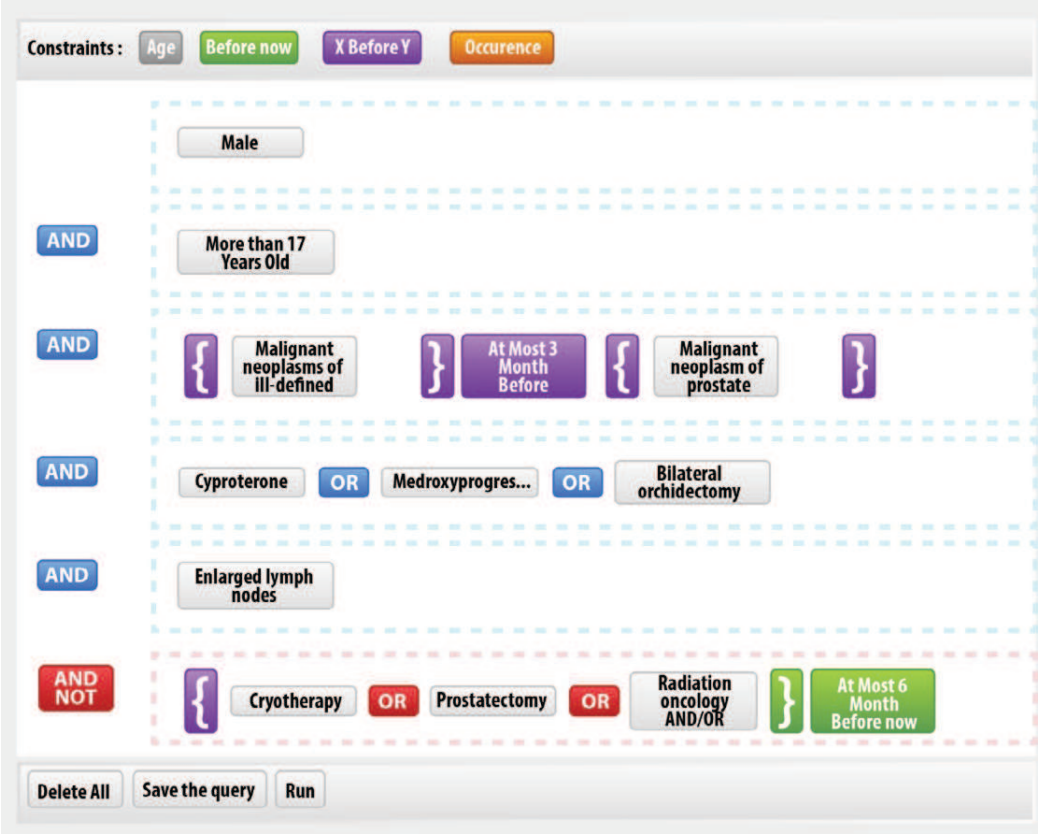


Figure 5: PFS query builder graphical user interface

After running an EC query, the results can be visualised by showing the overall results with the possiblity to access break-downs on the patient demographics (age categories and gender) level, the individual eligibility criterion level as well as the results returned by the individual sites (Figure 6).

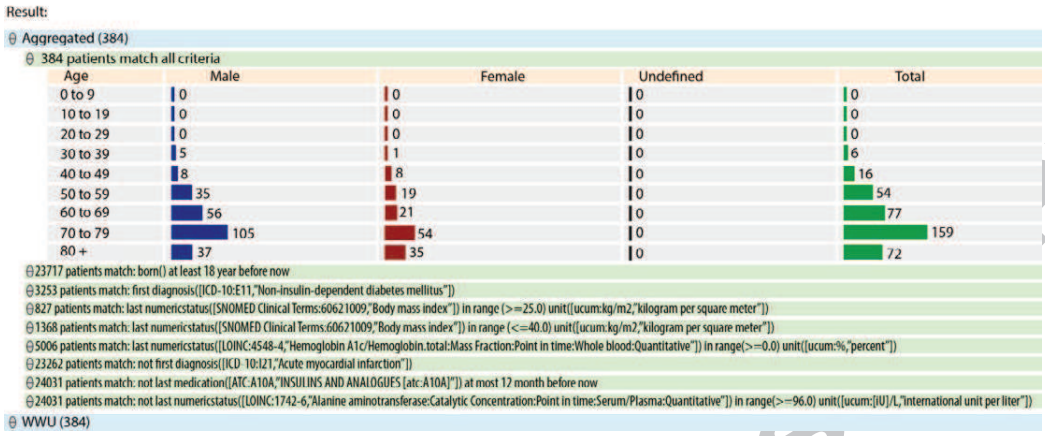


Figure 6: Screenshot of a query result

For the PRS this workbench has been extended to include recruitment study coordination functions and a corresponding dashboard showing the current recruitment and accrual status at each of the clinical sites that have been invited to participate in a given study.

The following figure illustrates the flow between the major PRS components (figure 7):

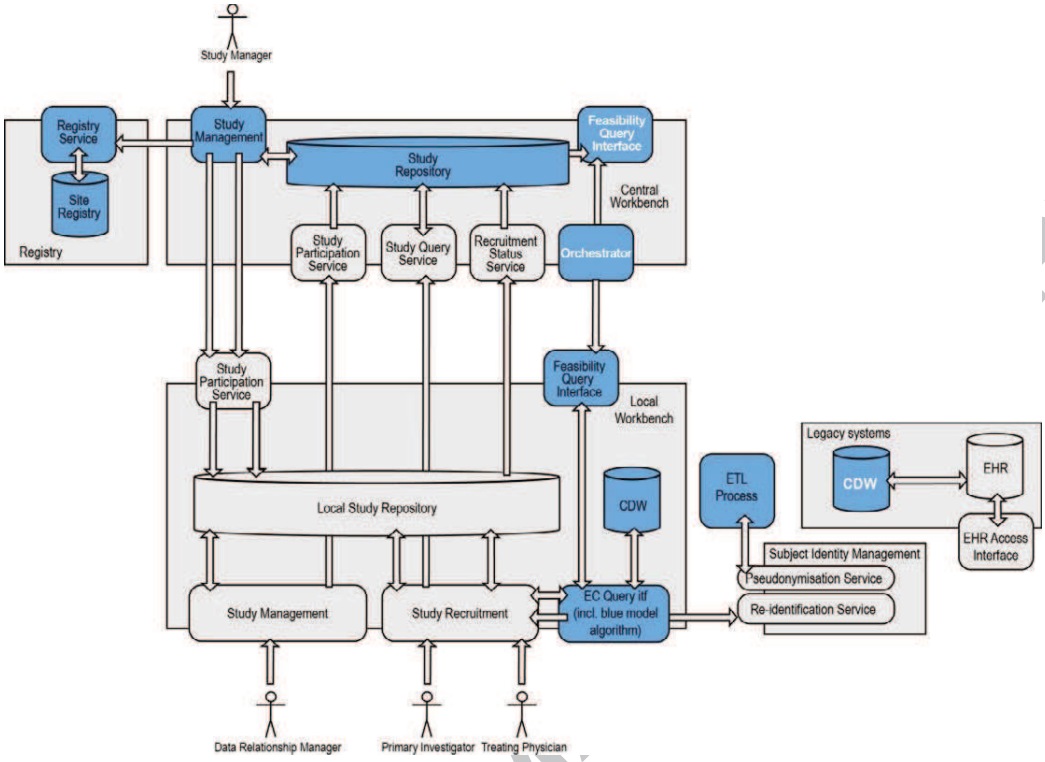
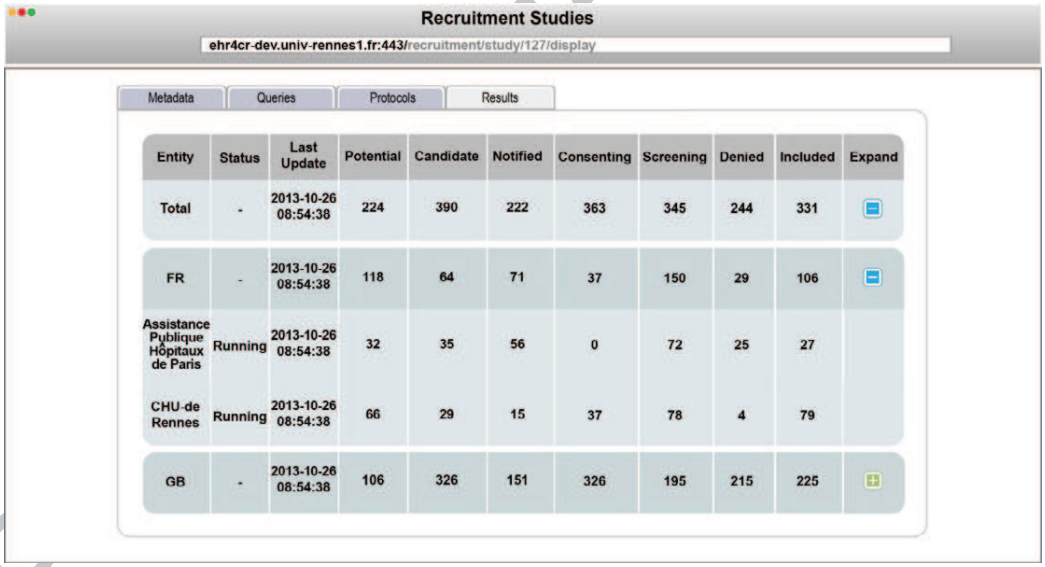


Figure 7: Flow between the PRS components (the dark blue components have been re-used from the PFS scenario and the light blue components have been added to support the PRS demonstration project)

In order to start the recruitment process for a given study, a new study definition must be created by the study manager. The definition includes the protocol description and optionally the formal eligibility criteria to allow computer-assisted checking of patient eligibility. The study definition can be based on an existing study definition previously created through the PFS or it can be newly created if protocol feasibility checking for the given study has not been previously conducted on the platform. The study definition can also be based on an existing CDISC SDM (Study Design Model) file. The formal eligibility criteria defined for the PFS can be extended and enhanced to be used for the PRS.

Through the registry service, the study manager is able to select clinical sites of interest that expose the necessary technical interface. Following this an invitation containing the study definition will be sent to each of the selected sites. The study definition will be imported in the local study repository and the invitation will eventually be presented to the *data relationship manager* responsible for engaging the clinical site in semi-automated studies.

Once a clinical site has been invited to participate in a given study for recruitment, its participation status will be visible to the study manager. Once the clinical site accepts to participate, the number of patients in each of the various recruitment stages will be periodically made available to the study manager (Figure 8).



Entity	Status	Last Update	Potential	Candidate	Notified	Consenting	Screening	Denied	Included	Expand
Total	-	2013-10-26 08:54:38	224	390	222	363	345	244	331	
FR	-	2013-10-26 08:54:38	118	64	71	37	150	29	106	
Assistance Publique Hôpitaux de Paris	Running	2013-10-26 08:54:38	32	35	56	0	72	25	27	
CHU de Rennes	Running	2013-10-26 08:54:38	66	29	15	37	78	4	79	
GB	-	2013-10-26 08:54:38	106	326	151	326	195	215	225	

Figure 8: Screenshot of the clinical site recruitment status dashboard

For the clinical sites, an entirely new application has been designed and implemented to support data relationship management (participations in clinical studies), local study

management (user assignments and study status) and candidate patient identification and patient recruitment status tracking. After the site accepts to participate in a given study, the Principal Investigator can create a selection containing potential candidate patients to be recruited. If the study participation request includes a formal representation of the EC, these can be used at the clinical site to automatically query the data access endpoint to populate the initial list of (potential) candidate patients (computer-assisted candidate selection). The initial candidate patient list will be based on pseudonimised records and patient identifiying information will not be visible until a treating physician has contacted the patient and if the patient agrees to enter the enrollment process.

The overall distributed workflow between study manager (central workbench) and the users at the clinical site (local workbench) is depicted in the following diagram (figure 9):

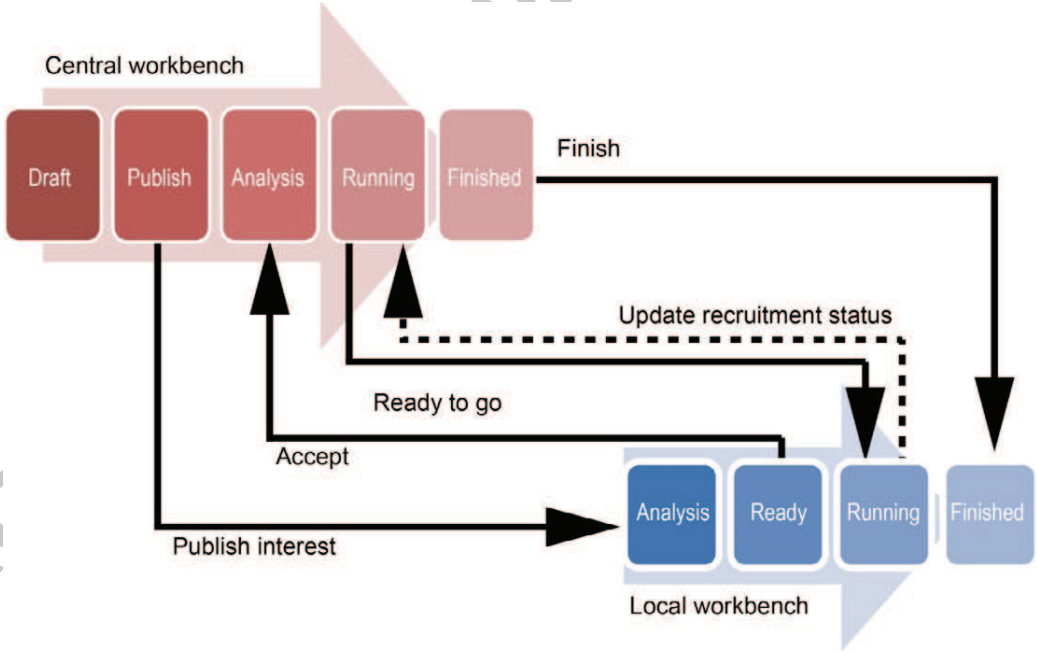


Figure 9: Overall distributed workflow

4.2.6 Security architecture

The EHR4CR security architecture is an important part of the overall architecture. It specifies a set of profiles that must be adhered to by EHR4CR-compliant tools and services in order to securely engage with one another. These profiles have been defined by selecting a number of existing security standards and by defining a number of restrictions on their use to achieve interoperability. Next to this, the EHR4CR security architecture also mandates the adoption of a number of information security practices (e.g. regarding the treatment of passwords and personal medical data) as mandated by the EHR4CR non-functional requirements. The following table mentions the most important security standards on which the EHR4CR security architecture is based:

Standard	**URL**	**Use**
SOAP	http://www.w3.org/TR/soap/	Required for all secure web service interactions.
WS-Security v1.1	https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wss	Required for achieving secured messaging between EHR4CR service providers.
WS-Trust v1.4	http://docs.oasis-open.org/ws-sx/ws-trust/v1.4/errata01/ws-trust-1.4-errata01-complete.html	Obtaining security tokens from a platform-governed security service to allow establishing, assessing and brokering trust relationships between EHR4CR service providers.
SAML v2	http://saml.xml.org	SAML web SSO profile for end-user authentication. Use of SAML v2 tokens for identity federation and platform-level authorization.
SAML V2.0	http://docs.oasis-	Mechanism for specifying delegation credentials.

Condition for Delegation Restriction	open.org/security/saml/Post2.0/sstc-saml-delegation-cs-01.pdf	
SAML 2.0 Profile of XACML, version 2.0	http://docs.oasis-open.org/xacml/3.0/xacml-profile-saml2.0-v2-spec-cs-01-en.pdf	Mechanism for requesting authorization decisions from a platform-governed authorization service and for requesting attribute assertions from Attribute Authorities in order to evaluate access control policies.

Table 2: Major standards on which the EHR4CR security architecture is based

This security architecture was chosen over other solutions given the maturity of the standards on which it is based, the level of software industry adoption, the availability of tools and development libraries, the degree of compatibility of these standards with the overall EHR4CR architecture and the familiarity of the EHR4CR security work package team with the above standards.

The EHR4CR security architecture has been selected to provide a number of state-of-the-art capabilities to deal with challenging security concerns. This includes the ability to enforce (potentially site specific) access control policies against the identity and characteristics (e.g. affiliation or geographic location) of the end-user on whose behalf EC queries have been issued even though the end-user is not invoking the data endpoints directly. In order to achieve this, the EHR4CR authentication and authorisation services support standard-based constrained delegation of credentials [41, 42].

In order to encompass existing local data provider security policies and firewall rules, the platform supports the invocation of web services using dynamically configurable transport bindings. Examples include asynchronous web service invocation by employing message-oriented middleware to provide the ability (from an endpoint provider perspective) to retrieve (*pull*) incoming queries rather than receiving these directly (*push*), thus avoiding the need for endpoint providers to accept incoming connections from the Internet into their local network. This feature ensures compliance with local data provider policies, thus facilitating platform adoption, while at the same time allowing for standards-based authentication and authorisation of end-users and web service clients operating on their behalf (Figure 10).

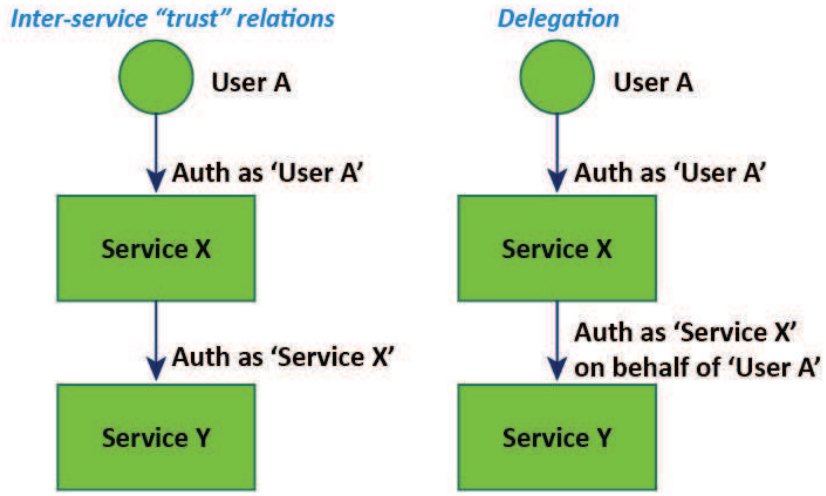


Figure 10: Classical "trust"-based inter-service authentication vs. delegation-based authentication

4.3 Pilots

A top list of data elements containing 75 EHR data elements has been identified by comparing common eligibility criteria used by EFPIA partners at the pilot sites with available data

elements in the EHR/CDW and EDC systems. In addition, a wish-list has been drafted of a further 21 data elements which were not available at more than 50% of the sites but deemed important [43].

Currently, all pilot sites have installed a local endpoint with connection to a local clinical data warehouse, and eleven data provider sites in five countries are connected to the EHR4CR platform.

With regard to the protocol feasibility scenario, the proof-of-concept demonstrator has been tested using feasibility queries from twelve different clinical trials. All EFPIA partners participated in this user acceptance test. Overall, 373 free-text eligibility criteria were reviewed by clinical trial experts. 175 feasibility criteria were transformed into a computable representation. Pilot sites mapped approximately 300 codes from their local terminologies into the central EHR4CR terminology, for instance taking into account different national coding systems for medical procedures.

Evaluation of PFS will compare the number of patients counts reported using traditional feasibility methods vs. patient counts obtained through EHR4CR platform, vs. manual count of eligible patients obtained through manual review of patient files.

With regard to the patient recruitment and identification scenario, fifteen trials have been identified by the EFPIA partners. Work has been undertaken in order to simplify their eligibility criteria. Evaluation of the PRS will take place by comparing patient lists and counts by two distinct methods (obtained through traditional recruitment methods vs. obtained through searching and identification of the EHR4CR platform).

5. Conclusion

This paper has described the rationale and methodology of the EHR4CR project and provided a detailed description of the first two services that have been implemented: for protocol feasibility and for patient identification for recruitment. The consortium is now working to implement further scenarios for linking EHR data to the EDC systems used to conduct clinical trials, to reduce the duplication and errors when pre-existing clinical information is re-entered during a trial and also for using this link to help better detect serious adverse events to trial medication. These service specifications will be published later.

Through a combination of a consortium that brings collectively many years of experience from previous relevant EU projects and the global conduct of clinical trials, an approach to ethics that engages many important stakeholders across Europe to ensure acceptability, a robust iterative design methodology for the platform services that is anchored on requirements and an underlying Service Oriented Architecture that has been designed to be scalable and adaptable, EHR4CR could be well placed to deliver a sound, useful and well accepted pan-European solution for the reuse of hospital EHR information to support clinical research studies. Consequently, EHR4CR could play a role in the discovery of new knowledge from health data, as well stimulate the ability of hospitals to promote excellence in documenting patients records in their EHRs.

By the end of this project there are clear expectations that the EFPIA companies and hospital partners involved in EHR4CR will be the first to become accredited members of the EHR4CR network. By making the EHR4CR project deliverables and specifications publicly available, as well as through a non-profit European Institute for Innovation through Health Data, the intent is to attract new actors to join this network. This may include in the future other providers of EHR4CR services and possibly alternative implementations of the platform itself, although

these will also be required to be certified according to a common conformance standard. The organisational model, with inclusion of an independent Trusted Third Party (TTP), will also allow for additional kinds of data transactions between different stakeholders and environments (including e.g. platform-level audit trail (re)construction and specific (de-identified) data exchanges outside the scope of the standard scenarios). Through the envisaged European Institute for Innovation through Health Data, pharmaceutical companies will be encouraged to continue to collaborate pre-competitively to evolve the EHR4CR services to meet new needs, and to accelerate and improve the quality of clinical research.

The project could enable a scalable and flexible approach to reusing EHR data which can bring safe and effective innovative medicines more quickly to the market and enrich medical research and knowledge. Encouragingly, top-down change is happening in tandem with bottom-up capability (i.e. EHR systems and the richness of EHRs are improving so that EHR4CR like solutions can succeed). Where regulatory authorities have historically been slow to evolve, a more welcoming approach to innovation is on the cards. As these approaches take shape, the stage is set for an efficient clinical research experience: one in which the design is better suited to a field undergoing constant transformations and shake-ups. For the industry, this will provide innovative integrated and efficient solutions. For patients, it will reveal to be a genuine revolution.

Acknowledgements

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking (http://www.imi.europa.eu) under grant agreement n° 115189, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

Conflict of interest statement

No conflict of interest was declared.

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012; 13(6): 395-405.
2. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, Dugas M, Dupont D, Schmidt A, Singleton P, De Moor G, Kalra D. Electronic health records: new opportunities for clinical research. J Intern Med 2013; 274(6): 547-560.
3. The added value of electronic health records (EHRs): Linking patient care, clinical research and public health, 2008. (http://www.eurorec.org/files/filesPublic/High_Level_Statement_EHR%20in%20EU__FINAL%20_July%2014.pdf).
4. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009; 48(1): 38-44.
5. Turisco F, Keogh D, Stubbs C, Glaser J, Crowley Jr WF. Current status of integrating information technologies into the clinical research enterprise within US academic health centers: strategic value and opportunities for investment. J Investig Med 2005; 53(8): 425-433.
6. De Moor G: EHR-Certification, semantic interoperability and the link to clinical research. eHealth-WoHIT Conference, 2010 (http://www.eurorec.org/news_events/newsArchive.cfm?newsID=216).
7. Dugas M, Lange M, Muller-Tidow C, Kirchhof P, Prokosch H-U. Routine data from hospital information systems can support patient recruitment for clinical studies. Clin Trials 2010; 7(2): 183–189.
8. Embi PJ, Jain A, Clark J, Harris CM. Development of an Electronic Health Record-based Clinical Trial Alert System to Enhance Recruitment at the Point of Care. AMIA Annu Symp Proc 2005; 231–235.

9. Séroussi B, Bouaud J. Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials. Artif Intell Med 2003; 29(1-2): 153–167.
10. National Institute of Health, National Center for Research Resources: Electronic Health Records Overview, 2006 (http://www.ncrr.nih.gov/publications/informatics/ehr.pdf).
11. Fineberg HV. A Successful and Sustainable Health System— How to Get There from Here? N Engl J Med 2012; 366(11): 1020-7.
12. Sundgren M, Wilson P, deZegher I. Making the most of the Electronic Age. *European Pharmaceutical Contractor* 2009; 3:18-21.
13. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, Leese A, Richardson C, Mantas J, Murray P, De Moor G. Trustworthy Reuse of Health Data: A Transnational Perspective. Int J Med Inform 2013; 82(1): 1-9.
14. PriceWaterhouseCoopers. Transforming healthcare through secondary use of health data, 2009. http://www.pwc.com/us/en/healthcare/publications/secondary-health-data.jhtml
15. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013; 20(1): 144-51.
16. Holzer K, Gall W. Utilizing IHE-based Electronic Health Record Systems for Secondary Use. Methods Inf Med 2011; 50: 319-325.
17. Cruz-Correia R, Rodrigues P, Freitas A, Almeida F, Chen R, A. CP. Data Quality and Integration Issues in Electronic Health Records. In: Hristidis V, editor. Information discovery on electronic health records. London: CRC Press 2010; 55-95.

18. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform 2012; 45(4): 763-771.
19. EHR4CR project website. http://www.ehr4cr.eu
20. El Fadly A, Rance B, Lucas M, Mead C, Chatellier B, Lastic PY, Jaulent MC, Daniel C. Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. J Biomed Inform 2011; 44 Suppl 1: S94-102
21. Ouagne D, Hussain S, Sadou E, Jaulent MC, Daniel C. The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. Stud Health Technol Inform. 2012; 180: 634-538.
22. Dipak Kalra, Andreas Schmidt, HWW Potts, Danielle Dupont, M Sundgren and Georges De Moor, on behalf of the EHR4CR Research Consortium. Case Report from the EHR4CR Project—A European Survey on Electronic Health Records Systems for Clinical Research. iHealth Connections, 2011;1(2):108–13.
23. NorthWest EHealth. Feasibility And Recruitment System for Improving Trial Efficiency (FARSITE). http://www.nweh.org.uk/farsite. Accessed on June 5th 2014.
24. Peterson KA, Fontaine P, Speedie S. The Electronic Primary Care Research Network (ePCRN): a new era in practice-based research. J Am Board Fam Med. 2006 Jan-Feb; 19(1): 93-7.
25. National Center for Biomedical Computing. Informatics for Integrating Biology & the Bedside (i2b2). https://www.i2b2.org . Accessed on June 5th 2014.
26. National Center for Biomedical Computing. Data Sharing Network (SHRINE). https://www.i2b2.org/work/shrine.html. Last accessed on June 5th 2014.

27. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20(e1):e147‑154.
28. Klann JG, Murphy SN. Computing health quality measures using Informatics for Integrating Biology and the Bedside. J Med Internet Res. 2013;15(4):e75.
29. Pathak J, Wang J, Kashyap S, et al. (2011) Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Informatics Assoc Jamia 18:376–386. doi: 10.1136/amiajnl-2010-000061
30. Boyce RD, Ryan PB, Norén GN, Schuemie MJ, Reich C, Duke J, Tatonetti NP, Trifirò G, Harpaz R, Overhage JM, Hartzema AG, Khayter M, Voss EA, Lambert CG, Huser V, Dumontier M. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. Drug Saf. 2014 Aug;37(8):557-67
31. The Office of the National Coordinator (ONC) Standards and Interoperability (S&I) Framework. http://www.siframework.org. Last accessed on September 15th 2014.
32. The Office of the National Coordinator (ONC) Standards and Interoperability (S&I) Framework – Query Health pilot. http://pilots.siframework.org/pilots.html. Last accessed on September 15th 2014.
33. Klann JG, Buck MD, Brown J, Hadley M, Elmore R, Weber GM, Murphy SN. Query Health: standards-based, cross-platform population health surveillance. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):650-6
34. Weng C, Tu SW, Sim I, Richesson R (2010) Formal representation of eligibility criteria: a literature review. J Biomed Inform 43:451–467. doi: 10.1016/j.jbi.2009.12.004
35. ISO/IEC 11179: Information Technology -- Metadata registries (MDR). http://metadata-standards.org/11179. Last accessed on September 15th 2014.

36. ISO/IEC 21090:2011. Health Informatics – Harmonized data types for information interchange.

37. Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: a technical report. J Am Med Inform Assoc 2008; 15: 130-7.

38. Observational Medical Outcomes Partnership (OMOP) Common Data Model. http://omop.org/CDM. Last accessed on September 15th 2014.

39. The EHR4CR semantic models, resources & semantic tools. http://termapp.limics.fr/ehr4cr. Last accessed on September 15th 2014.

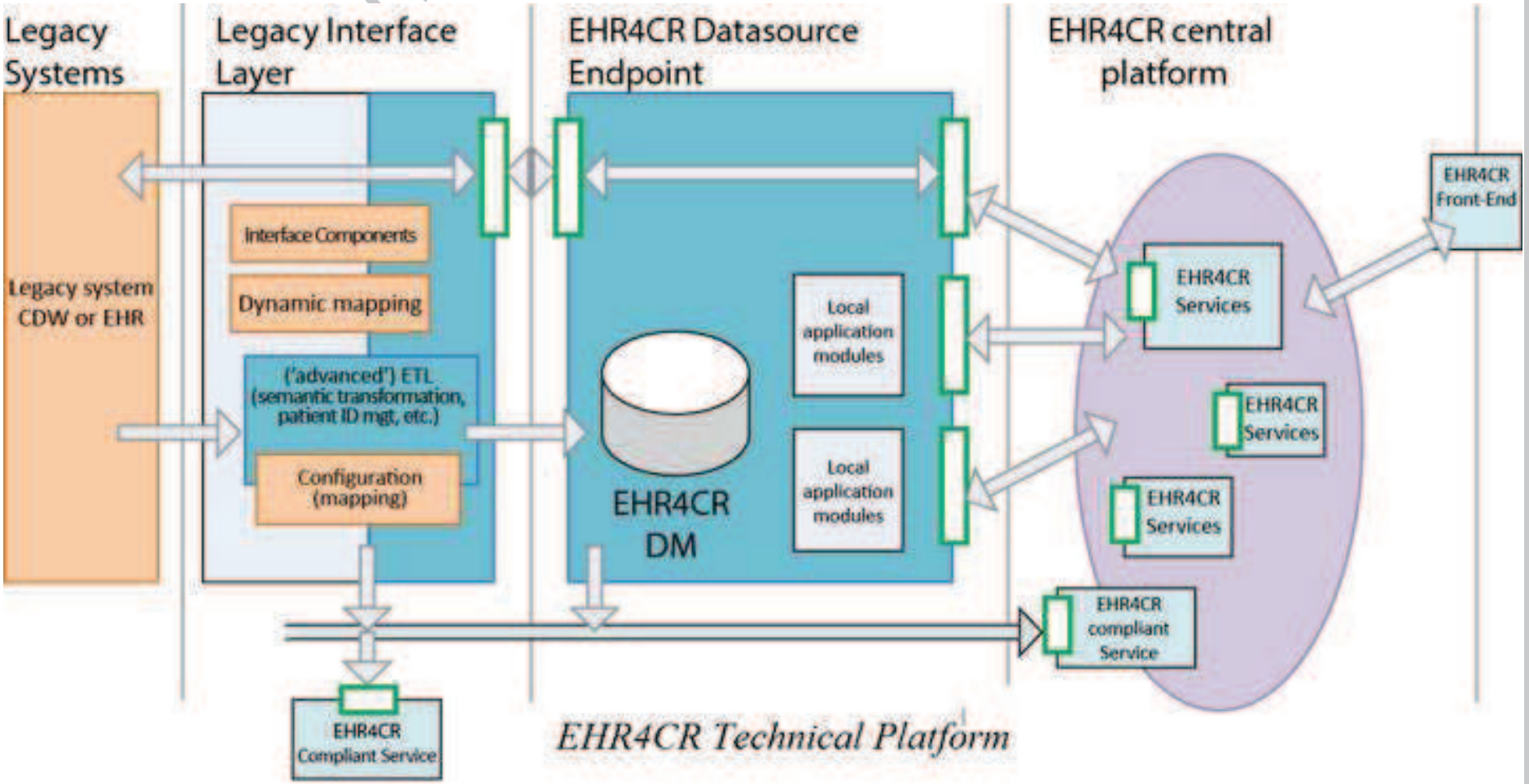
40. Bache R, Taweel A, Miles S, Delaney BC. An eligibility criteria query language for heterogenous data warehouses. Methods Inf Med 2014; 53(4)

41. OASIS Web Services Secure Exchange (WS-SX) TC. WS-Trust v1.4. Advancing Open Standards for the Information Society (OASIS). http://docs.oasis-open.org/ws-sx/ws-trust/v1.4/errata01/ws-trust-1.4-errata01-complete.html

42. OASIS Security Services TC. SAML v2.0 Condition for Delegation Restriction Version 1.0. Advancing Open Standards for the Information Society (OASIS). http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-delegation-cs-01.pdf

43. Doods J, Botteri F, Dugas M, Fritz F, EHR4CR WP7. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials 2014; 15:18.

*Graphical Abstracts (for review)



- EHR4CR has designed a secure platform to optimize clinical trial protocols
- The EHRCR platform facilitates patient identification and recruitment
- The EHR4CR platform uses state of the art security and has robust information governance policies
- The EHR4CR platform does not require the extraction or communication of any patient level data from hospitals
- Proof-of-concept demonstrators have been built and evaluated